37

/

APPROACHES ON AN EXPERIMENTAL CHINESE ELECTRONIC DICTIONARY

Shih-shyeng Tseng*, Meng-yuan Chang*, Ching-Chun Hsieh**, and Keh-jiann Chen**

* Computing Center, Academia Sinica,
Nankang, Taipei 11529, Taiwan, ROC
** Institute of Information Science, Academia Sinica,
Nankang, Taipei 11529, Taiwan, ROC

ABSTRACT

The CED, an experimental Chinese electronic dictionary, is essentially a particular textual database providing users with available information of about 40,000 Chinese vocabularies. For each vocabulary, the information includes a character string which is the vocabulary itself, its radical and stroke count of leading character, its pronounciations, its meaning descriptions, and its syntactic categories with annexed attributes. The CED consists of four portions : a set of MD files, a number of index files, a management subsystem, and a user interface. In addition to the traditional search functions on a Chinese dictionary, the CED also provides the wild-card function and the free term search function. The CED parser can construct the MD files from a source text efficiently and economically. In the paper, the organization of CED, the CED parser, and the retrieval mechanism of the CED will be presented.

1. INTRODUCTION

In Taiwan, research on Chinese information processing (abbr. CIP), including theoretical studies and application developments, has been proceeding for more than 17 years [1]. During the past years, most of the theoretical studies on CIP concerned Chinese characters, such as the studies and implementations of Chinese character sets [2,3], the coding techniques of Chinese characters [4,5], the input methods and output mechanisms of Chinese characters [6], etc. All the CIP systems based on Chinese character can only process the Chinese information as a number of character strings. Since words (vocabularies) are the smallest meaningful units of Chinese, we need more theoretical studies on Chinese vocabularies to develop a more intelligent CIP system. For that reason, the researchers of both the Computing Center of Academia Sinica and the Electronic Research and Service Organization have started a cooperative project to study the syntax and semantics of Chinese and intend to develop a parsing system for Chinese sentences. The development of the CED is basically intended for building a computerized tool to assist our research work.

Because the Gwoyeu Ryhbaw Tsyrdean is one of the most popular Chinese dictionary in Taiwan, and it collects a large volume of most-frequently-used Chinese vocabularies, we chose it as the source of our study. The CED currently contains the information of about 40,000 Chinese vocabularies. For each vocabulary, the information includes a character string which is the vocabulary itself, its radical and stroke count of leading character, its pronouciations, its meaning descriptions, its syntactic categories with annexed attributes. The first four are obtained from the Gwoyeu Ryhbaw Tsyrdean while the last one is obtained through a research work accomplished by our research team [7].

2. THE ORGANIZATION OF CED

The CED is composed of four portions : a set of machine-readable data files, called MD files, a number of index files, a management subsystem, and a user interface, as shown in Figure 1. The MD files form a context tree in order to provide a necessary search mechanism, such as that discussed in [8]. The context tree of MD files is shown in Figure 2. The information of one Chinese vocabulary, which is introduced in section 1, is stored up in a word record as a leaf context. Several word records are filled into a constant-size segment. A number of segments are then grouped into a MD file. In CED, there are about 40,000 word records stored up in twenty MD files, corresponding with the orderly vocabularies of Gwoyeu Ryhbow Tsyrdean. The index files will be discussed in Section 4.

The management subsystem consists of a set of maintenance programs and data-retrieving programs. They are actually a group of subroutines written in C language, in which each subroutine

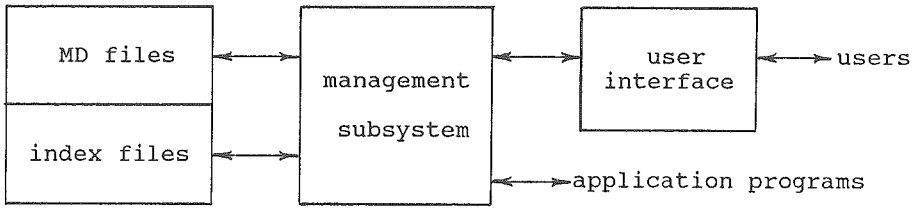


Figure 1 The organization of CED

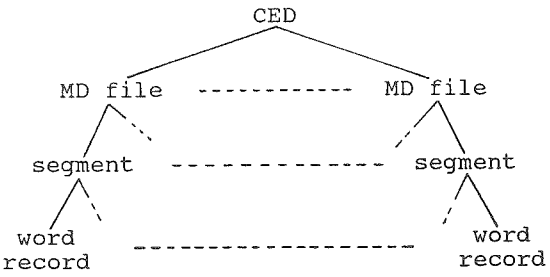


Figure 2 The context tree of MD files

performs a specified operation. The maintenance subroutines can be invoked by any application program to perform the operations of insertion, deletion and modification in MD files. The index files must be updated when some MD files are maintained. The data-retrieving subroutines can retrieve a group of word records corresponding to a given search expression. Any additional subroutines can be added to the management subsystem if necessary.

The last portion of CED is a user interface. It provides a user-friendly menu-driven mechanism to help users retrieve the information of Chinese vocabularies. It also provides the operations of insertion, deletion, and modification, but these operations just allowed for our researchers. In Section 5, we will go into a more detail about the user interface.

3.THE CREATION OF MD FILES

To develop the CED, the first step is to generate the MD files from the source text of the Gwoyeu Rybaw Tsyrdean. The MD files might be built by means of a traditional procedure. That is, a group of researchers laboriously analyzed each paragraph of the original dictionary to obtain a set of elements of the vocabulary, such as the radical and stroke count of leading character, pronounciations, and meaning descriptions. They also has to write those elements into a pre-printed form. Then the data in the filled forms should be typed into a computer system by a group of typists to create MD files. Since the Gwoyue Ryhbaw Tsyrdean contains more than 1.6 million Chinese character, the traditional method will lead a time-consuming and laborious work.

Instead of the traditional procedure an ingenious method had been applied to the creation of MD files. Principally, the loagical organization of the source text will be explicitly represented by its typsetting format. For example, Figure 3 shows the context structure [8] of Gwoyeu Ryhbaw Tsyrdean, and Figure 4 shows a piece of the dictionary. When a Chinese looks for a word in a Chinese dictionary, the context structure shown as Figure 3 will occur to his mind. The fact leads the idea that the text elements (e.g. elements described in previous paragraph) will be recognized by a program if the source texts with associated information of typesetting formats had been stored in computer files.

The new method which used to build the MD files on the source text is composed of two steps :

(1) The source text and associated information of typesetting format were entered into computer files by a group of typists. As an example, there are several different fonts of Chinese characters in Figure 4. In order to represent those different fonts, some markup symbols had been used to enclose the characters which appeared in each of text elements of radiceals, stroke counts and vocabularies, such as that discussed in [9].

(2) A program called CED parser scaned those source text files to recognize every text element and then to create the MD files. In order to recognize the text elements and then to construct each word record as result, a pattern-matching and solt-filling algorithm had been applied to the CED parser.

Since the operations of CED parser replaced the work of reserachers who analyzed the element and filled them into pre-printed form, the new method is more efficient and more economic than the traditional method. It took about three man-months to design and implement the CED parser.

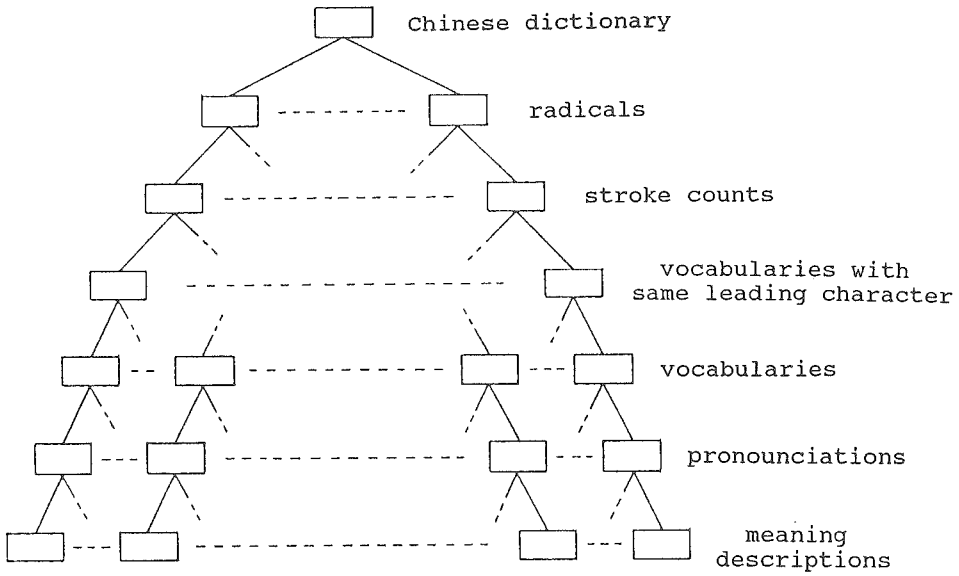


Figure 3 The content structure of traditional Chinese dictionary

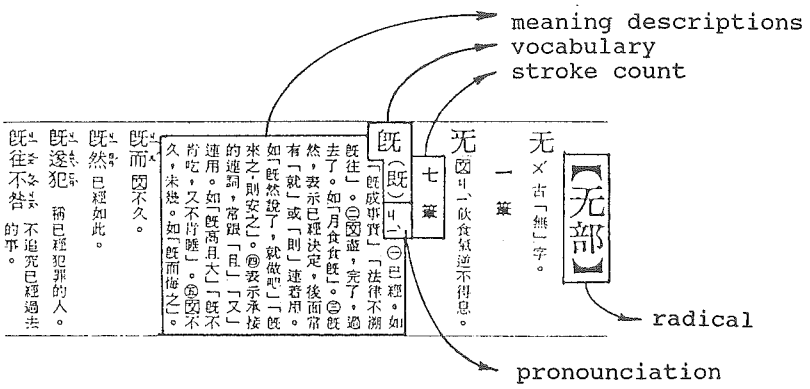
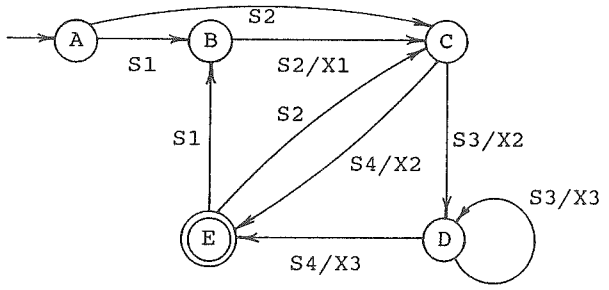


Figure 4 A sample piece of Gwoyeu Ryhbaw Tsyrdean

The pattern-matching and slot-filling algorithm can be simply illustrated by using the example shown in Figure 5. A pattern is essentially a list consisting of a series of alternatively appeared constants, such as S1, S2, S3 and S4, shown in Figure 5(a), and vaiables, such as X1, X2 and X3, also shown in Figure 5(a). The constants are the specific characters and/or symbols and the variables denote the pieces of text that we need. The frame cosists of a number of tags, such as T1, T2 and T3, shown in Figure 5(a), and slots, in which each tag is an additional symbol used to identify the variable follows it. Each slot of the frame must be given a name which previously appeared in a variable part of the pattern. In Figure 5(a), [....] denote an optional item and {....} denotes repetition of any times. The pattern-matching and slot-filling algorithm can be performed by a special parser, such as CED parser. The Figure 5(b) describes the operation of this parser. In Figure 5(b), an alphabet enclosed by a single circle denotes a state of the parser, an alphabet enclosed by double circles represents the final state of the parser. And an arc with label Si/Xj from state P to state Q denotes a state transition from P to Q with input Si and output Xj, i.e., the parser finds the constant Si from the source text and fills the variable Xj into the slot named by Xj. The parser write a filled frame to a destination file when the final state E is reached. The states B, C, D and E must be circulated many times until the source text has been completely scanned. During each circulation, if no S1 and <X1> are found in the source text, then the value of slot X1 will not be changed. And if more than one S3 and <X3> are scanned, then there are multiple X3's to be filled into the frame.

```
Pattern: [S1 <X1>] S2 <X2> {S3 <X3>} S4
Frame:   T1 <X1> T2 <X2> {T3 <X3>}
```

(a) the sample pattern and frame



(b) the state diagram for (a)

Figure 5 An example of pattern-matching and slot-filling algorithm

4. THE SEARCH FUNCTIONS AND INDEX STRUCTURES

The CED provides two types of search functions : the conventional search function and the free term search function. The conventional search function is that one searches the vocabularies by means of the content sturcture of a Chinese dictionary. That is, the radical and stroke count are given first to find a set of characters, and then from this set we choose one as the leading character that we need. A set of vocabularies in which each has the same leading character can be found by means of the selected character. Finally, we select the vocabulary that we need from the set of vocabularies. In order to provide the conventional search function, an index structure which is like the content sturcture shown in Figure 3 must be included in the index files.

The free term search functions can be performed by means of a refined character inversion method (abbr. ARCIM), such as that discussed in [8]. Thus a character inversion table (abbr. CIT) and a posting lists table (abbr. PLT), as shown in Figure 6, are also included in the index files. In order to provide the free term search functions, the information of a vocabulary is divided into four parts : the vocabulary itself, the pronounciations, the meaning descriptions, and the syntactic categories with annexed attributes. For the first two, just the wild-card function (i.e. partial matching) is allowed. The search expressions can be independently assigned to these four parts. After the search expressions are given, the vocabularies which satisfy these expressions are found by means of ARCIM.

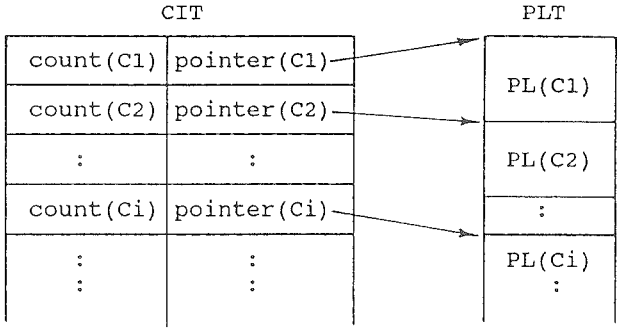


Figure 6 The index structure of ARCIM

5. THE USER INTERFACE

The CED essentially played the role of a computized tool to our researchers. So we need an interactive user interface to assist us to retrieve and sometimes to list out any information from MD files and moreover to maintain the contents of MD files if we want to. The user interface served a menu-driven operation were looked like it appeared in some packages of databases. The menu should be displayed in a terminal whenever the user interface activated. The user can choose one of the functions such as retrieval, listing, insertion, deletion and modification from the menu. Then the user interface responds and the user has to enter some data with the key board to accomplish this function step by step. Form examples, if we need to add a new syntactic category to a vocabulary, we first choose the function of modication from the menu, and then the user interface will ask which vocabulary we want. We have to locate the vocabulary we need by specifying some retrieving key(s). When the object appeared on the screen, we then identify the item of category. Following that, we select some descriptions of meanings which correspond to the new category and enter the new category with their annexed attributes. After those steps, the information of the new syntactic category should be added into the vocabulary.

The important constituent of the user interface is a screen controller. Under its control, the screen is divided into six windows. The first window displays the functions proviced by the user interface. The last window displays the system messages for current operation mode. The second window displays the vocabouary and the radical and the radical and stroke count of leading character. The pronounciations, meaning discriptions and syntactic categories with annexed attributes ard shown in the third, fourth and fifth windows, respectively. Since the third, fourth and fifth windows may be too small to display all contents, the screen controller provids a rolling function on those windows. In addition, the user interface can serve multiple users.

6. CONCLUSION

The CED has some potential applications, as dicussed following.

(1) A computerized dictionary which is deveoped for the users is principally a read-only database of vocabularies. Since the total storage space of CED can be limited to smaller than 10 million bytes and the user interface of CED includs a easy learn, easy use, and friendly retrieval mechanism, so the CED will be easily devdloped into a commercial product on some personal computers.

(2) Each of the information processing systems which are concerned with natural languages, such as machine translators, natural language query processors, expert systems, etc., has to include a dictionary. Since the CED contains 40,000 Chinese vocabularies, associated information, and a set of management programs, so it has the ability to provide the necessary basis for various Chinese information processing systems which are concerned with Chinese language.

An important feature of CED is that it is an open system. That is, any new information for some vocabularies and moreover any new vocabularies with associated information can be easily added to the CED. Reversely, a subset of vocabularies with optional a subset of information of those vocabularies can be easily abstracted from the CED to serve variuos applications.

In addition, the experience with the design and development of the CED is very useful to the development of information processing systems. For example, how to build up the data files of the textual database from a lage volume of source texts is an important but difficult problem. The experience in the project on the automation of Chinese History Literatures [9] which is another research project currently worked by researchers at Academia Sinica, as well as the experience in the development of CED had lead us to develop a new technique to solve this problem.

ACKNOWLEDGEMENT

Research of this paper was partially supported by Electronic Research and Service Organization, Industrial Technology Research Institute, Taiwan, R.O.C.

REFERENCE

1. Tseng, Shih-shyeng, *The Design and Implementation of the Chinese Character Characteristics Database*, Chinese Publication, Master Thesis, National Taiwan Institute of Technology, Taipei, Taiwan, June 1982.

2. Lin, Shuh, *A Statistical Study on Chinese Character Set for Computer Uses, NCTU Technical Report CC-601,* Chinese Publication, National Chiao Tung University, Hsinchu, Taiwan, March 1972.

3. The Chinese Character Analysis Group, *Symbol and Character Tables of CCCII,* Taipei, May 1983.

4. Hsieh, Ching-chun, et.al., "The Design and Application of the Chinese Character Code for Information Interchange (CCCII)", International Workshop on Chinese Library Automation, Taipei, Taiwan, Feb. 14-19, 1981.

5. Tseng, Shih-shyeng, et.al, "An Universal Coding System for Multi-lingual Environment", 52nd IFLA General Conference, Tokyo, Japen, Aug. 24-29, 1986.

6. Institute of Information Industry, *The Review and Analysis Report on Chinese Computers, Vol. 2, Technical Report C-18,* Chinese Publication, Taipei, July 1982.

7. Chang, Lily, et.al, *The Category Analysis on Chinese*, Chinese Publication, Computing Center, Academia Sinica, Taipei, July 1986.

8. Tseng, Shih-shyeng, Yang, Chen-chau, and Hsieh, Ching-chun, "The Document Representation and A Refined Character Inversion Method for Chinese Textual Databases", 1988 International Conference on Computer Processing of Chinese and Oriental Languages, Toronto, Canada, Aug. 29-Sept. 1, 1988.

9. Hsieh, Ching-chun, et.al., *The Design and Implementation on the Chinese Full-text Processing System*, Chinese Publication, Computing Center, Academia Sinica, Sept. 1986.